4

AD

Technical Note 6-89

COMPUTER PROGRAMS FOR MEASURING VARIATIONS IN SPEECH

Christopher C. Smyth

July 1989
AMCMS Code 612716.H700011

Approved for public release;
distribution is unlimited.

U.S. ARMY HUMAN ENGINEERING LABORATORY
Aberdeen Proving Ground, Maryland 21005-5001

®VAX 11/780 and DECtalk are registered trademarks of Digital Equipment Corporation.

Destroy this report when no longer needed.
Do not return it to the originator.

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Use of trade names in this report does not constitute an official endorsement or approval of the use of such commercial products.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION / DOWNGRADING SCHEDULE

3 DISTRIBUTION / AVAILABILITY OF REPORT
Approved for public release;
distribution is unlimited.

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
Technical Note 6-89

5. MONITORING ORGANIZATION REPORT NUMBER(S)

6a. NAME OF PERFORMING ORGANIZATION
Human Engineering Laboratory

6b. OFFICE SYMBOL *(If applicable)*
SLCHE

7a. NAME OF MONITORING ORGANIZATION

6c. ADDRESS *(City, State, and ZIP Code)*
Aberdeen Proving Ground, MD 21005-5001

7b. ADDRESS *(City, State, and ZIP Code)*

8a. NAME OF FUNDING / SPONSORING ORGANIZATION

8b. OFFICE SYMBOL *(If applicable)*

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS *(City, State, and ZIP Code)*

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
6.27.16	1L162716AH70		

11. TITLE *(Include Security Classification)*
Computer Programs for Measuring Variations in Speech

12. PERSONAL AUTHOR(S)
Christopher C. Smyth

13a. TYPE OF REPORT
Final

13b. TIME COVERED
FROM ________ TO ______

14. DATE OF REPORT *(Year, Month, Day)*
1989, July

15. PAGE COUNT
26

16. SUPPLEMENTARY NOTATION

17. COSATI CODES

FIELD	GROUP	SUB-GROUP
23	02	
25	04	

18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
speech analysis
computer programs
speech variations

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*

The measurement of variations in speech is investigated. Computer programs developed to study speech patterns determine the word boundaries and spacing, the fundamental pitch and the formants of words, and classification of the speech as voiced or unvoiced.

The potential use of the computer programs is demonstrated using synthetic speech. A synthetic voice generator was used to generate voice patterns from a known "speaker" in two different voice patterns by changing "pitch" and "breathing" parameters. The patterns were digitized and processed on a VAX® 11/780 computer for analysis. The computed parameters agree with those used to generate the synthetic speech.

It is planned to use these programs in future investigations of the relation between work load stress and voice variations in human factors testing of subjects who perform data entry tasks using automatic voice recognition systems.

20. DISTRIBUTION / AVAILABILITY OF ABSTRACT
☒ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Doris S. Eanes

22b. TELEPHONE *(Include Area Code)*
(301) 278-4478

22c. OFFICE SYMBOL
SLCHE-SS-IR

Previous editions are obsolete.

AMCMS Code 612716.H700011 Technical Note 6-89

COMPUTER PROGRAMS FOR MEASURING VARIATIONS IN SPEECH

Christopher C. Smyth

July 1989

Approved: John D. Weisz
JOHN D. WEISZ
Director
Human Engineering Laboratory

Approved for public release;
distribution is unlimitec

U.S. ARMY HUMAN ENGINEERING LABORATORY
Aberdeen Proving Ground, Maryland 21005-5001

LIST OF TERMS, ABBREVIATIONS, AND ACRONYMS

auto-correlation function	a measure of the dependence of future signal values upon the past values
breath group	a group of sounds uttered in one breath
cepstrum	the inverse Fourier transform of the logarithm of the power spectrum of a signal
formant	resonance frequencies of the vocal tract when producing a vowel sound
Fourier transform	a mapping of a signal from a function of time to a function of spectral frequency; the inverse maps from the frequency domain to the time domain
fricative	pronounced by forcing the breath, either voiced or unvoiced, through a narrow slit formed at some point in the mouth
Hamming window	a set of numerical values used in spectrum analysis to multiply the samples from a signal segment with the purpose of decreasing the influence on the spectrum of the signal values at the segment end points
lexical	describing words as isolated items of vocabulary rather than as elements in a grammatical structure
Linear Predicitive Coding (LPC)	a parametric, nonlinear method of describing a time varying signal as the output of a linear filter; the method employs auto-correlation functions to compute the coefficients of the filter
maxima over time	maximum value of intensity measured over a specified period of time
sonorant	a voiced consonant that is less resonant than a vowel but more so than an unvoiced plosive
spectrum	the intensity of sound arranged by the order of the respective wavelengths

Virtual Memory System (VMS) — a software operating system for the DEC VAX® 11/780 computer

zero crossing — the occurrence where a signal changes its value from positive to negative or negative to positive

CONTENTS

APPENDIX

FIGURES

TABLES

COMPUTER PROGRAMS FOR MEASURING VARIATIONS IN SPEECH

INTRODUCTION

Measurement of the variations in speech patterns induced by task stress on subjects is of interest in the field of human factors testing of military systems, especially in the area of automatic speech recognizers used as voice data entry devices. A knowledge of the voice variations may explain decreases in recognizer performance during test scenarios. A voice data entry task using a template-matching connected word automatic speech recognizer is performed with speech entries of a fixed format for each part of the task. These fixed formats of speech are entered by the subject in an enrollment stage before use with the speaker-dependent recognizer. Variations from these enrollment patterns during testing may cause misrecognitions or nonrecognitions. A determination of stress points for subjects in a test scenario would provide intrinsic measures of the complexity of military equipment and tasks and the suitability of using a voice recognizer in those tasks.

There is evidence that stress can alter the voice of a subject enough to influence the accuracy of an automatic voice recognition system. French (1983) and Poock and Martin (1984) (also Martin and Poock, 1984) have investigated the effects of emotional and perceptual motor stress in subjects on the accuracy of voice recognition devices. Their conclusion was that task stress greater than that induced in subjects during the enrollment phase can cause a decrease in the recognition accuracy. It therefore follows that the decrease in recognition accuracy is caused by a change in the subject's voice from the enrollment pattern, which is induced by the task stress. The measurement of the voice variations may bear a correlation to the task stress.

The general approach taken was to review the literature to ascertain the effects of stressful or emotional situations on various voice parameters during previous research. Computer programs were developed to measure some of the primary features of the speech signal. To test the effectiveness of the computer programs, a demonstration was arranged. Two voices, which were consistent with the characteristics indicated in the literature, were programmed into a DECtalk® synthetic speech generator. The DECtalk® output was then manipulated by the computer programs to determine whether the analysis was consistent with previous research.

OBJECTIVES

The objectives of this report are to review the literature about the effects of stress on speech, to describe computer programs developed for measuring voice variations in speech, and to discuss their potential employment.

REVIEW OF THE LITERATURE

Most of the available literature discusses the effects of emotional stress on speech. The authors reported the results of experiments conducted with subjects in both nonemotional and emotional situations, actors who simulated the corresponding situations, or real life recordings. The parameters ascertained to be most related to speech variations induced by stress are (1) fundamental pitch frequency, (2) variation in pitch, (3) shift in formants, (4) rate of speech deliverance, and (5) interword spacing. These parameters were ascertained to be interrelated.

Hicks (1979) studied the effect of stress on the fundamental pitch frequency and the speech-to-pause ratio. He reported that the fundamental pitch frequency of speech is influenced by stress, and that this parameter is equally likely to go up or down from the unstressed speech values, with concomitant frequency range variations. The magnitude and direction of the frequency shift is a function of the individual speaker. Hicks reported that the ratio of speaking time to total time increased for stressed speakers. Stressed speakers talked in longer bursts with shorter pauses separating the bursts. This increase in speech-to-pause ratio correlated positively with stress.

Williams and Stevens (1972) reported that the emotions of anger, fear, and sorrow produce characteristic differences in the contour of the fundamental pitch frequency, the average speech spectrum, temporal characteristics, the precision of articulation, and waveform regularity of successive glottal pulses in stressed speech. They studied the correlation between emotions and speech by employing professional method actors. By analyzing high quality recordings of the actors reading emotional dialogue, the study attempted to identify and measure those parameters of speech that reflect the emotional state of the speaker.

The authors reported that the aspect of the speech signal that appears to provide the clearest indication of the emotional state of a talker is the contour of the fundamental pitch frequency over time. This contour has a stereotype shape for a breath group that is generated in normal speech without marked emotions of any kind. The normal contour is characterized by smooth, slow, and continuous changes in the fundamental pitch frequency as a function of time, with the changes occurring in syllables on which emphasis or linguistic stress is to be placed. Emotions appear to have several effects on this basic contour shape.

The spectrograms for neutral voicing show a well-defined structure during the speaking of vowels, with little noise or irregularities either between the formants or in the high frequency range where formants are often not visible. Consonants are frequently uttered in an imprecise manner, particularly in unstressed syllables. Sentences are usually generated with shorter durations during unstressed situations than for emotional situations.

The effect of the emotion, anger, on speech is a high value of the fundamental pitch frequency which persists throughout a breath group. This increase is, on the average, at least half an octave above the fundamental pitch frequency for a neutral voice. The range in variation of fundamental pitch frequency is also considerably greater than the range for the neutral voice free of emotion and stress. Some syllables are produced with increased intensity or emphasis, and the vowels in these syllables have the highest fundamental pitch frequency. Syllables spoken with increased intensity or emphasis tend to have weak first formants, and are generated with some voicing irregularities due to irregular fluctuations from one glottal pulse to the next. The basic opening and closing articulatory gestures characteristic of the vowel-consonant alternation in speech appear to be more extreme when a speaker is angry. The vowels tend to be produced with a more open vocal tract, and hence have a higher first formant frequency. The consonants are generated with a more clearly defined closure. The durations of utterances spoken in anger are usually longer and the syllabic rate lower; however, this effect is not great and is not always consistent for all voices.

The average fundamental pitch frequency for the emotion of fear is lower than that observed for anger, but higher than that for the neutral tone, although for some voices, it is close to that for a neutral voice. Occasionally, however, peaks in the fundamental pitch frequency occur that are much higher than those encountered in a neutral situation. These peaks are interspersed with regions where the fundamental pitch frequency is in a normal range. The pitch contours in the vicinity of the peaks sometimes have unusual shapes, that is, irregular bumps or discontinuities, and voicing irregularities are sometimes present. The duration of an utterance tends to be longer in the fearful state than in the anger or neutral situation. The vowels and consonants are more precisely articulated than in a neutral situation. Observations of spectrograms, however, reveal no clear and consistent correlation.

The average fundamental pitch frequency for the emotion of sorrow is considerably lower than for neutral situations, and the range of variation in the fundamental pitch frequency is quite narrow. This change in fundamental pitch frequency is accompanied by a marked decrease in the rate of articulation and an increase in the duration of an utterance. The syllabic rate, on the average, is slowed down by at least half from that of a neutral voice. The increased duration results from longer vowel and consonant utterances and from pauses that are often inserted in a sentence. Perhaps the most striking effect is voicing irregularities. On occasion, the voiced sounds will reduce to noise level due to whispering.

Williams and Stevens (1972) concluded that measurements of the median fundamental pitch frequency and the range for a sample of speech of several seconds' duration may classify a speaker's emotional state. Assuming that the normal pitch and range are known for a talker, a reduced pitch median and range may classify the talker's emotional state as one of sorrow, while an increased pitch median and range may indicate an angry or fearful state. Additional information is given by certain attributes of the pitch contour, shifts in spectrum, changes in duration, and voicing irregularities.

A study was made by Lieberman (1961) of the rapid fluctuations that occur in the fundamental excitation pitch of speech in certain emotional modes: boredom, confidential, doubt, fear, happiness, question, statement, and pompous, for six male speakers of the American English. Comparing the durations of adjacent pitch periods for the different modes, Lieberman determined that the durations did not remain constant, but that the magnitude of the differences increased as the duration approached 6 milliseconds (ms). Beyond that, there was no consistent trend. Somewhat greater differences between the durations of successive periods occurred in those samples assciated with the onset and end of voicing and suddpn spectral shifts. Only small differences occurred with those emotional modes that seemed to requ:re greater conscious vocal control in their production.

Lieberman and Michaels (1962) investigated the relationship of fundamental pitch frequency and speech envelope amplitude to the emotional content of speech for three male speakers of the American English who spoke in the same emotional modes examined above. They extracted the pitch and speech envelope amplitude and fed combinations of these signals into a synthesizer. The pitch perturbations were smoothed, and the output amplitude was modulated with the envelope amplitude. The results were categorized by naive listeners for emotional modes. The results showed that pitch alone was not sufficient for mode determination. While the listeners were able to correctly identify 85 percent of the emotional content of unprocessed speech, the pitch information alone resulted in correct identification only 44 percent of the time. Adding amplitude information to the pitch increased the correct identification to 47 percent. Smoothing the pitch information reduced the identification. Smoothing with a 40-ms time constant reduced the correct identifications to 38 percent, while smoothing with a 100-ms time constant resulted in 25 percent correct identification.

Levin and Lord (1975) investigated fundamental pitch frequency as an indicator of emotional state and the use of cepstral analysis to determine these changes. They performed measurement analyses in two octave ranges: the first included primary pitch components, and the second included altered pitch components and harmonics. The selection of ranges was based on the population modal values of fundamental pitch frequency: 120 Hz for males and 220 Hz for females. The measurement ranges for male speakers were 80 to 160 Hz and 160 to 320 Hz. The ranges for females were 150 to 300 Hz and 300 to 600 Hz. The ranges represented contiguous octaves such that differences in the range-averaged pitch would be indicative of emotional change. The extraction of the pitch frequencies from the speech signal was done by cepstral analysis which is used to separate the vocal tract effects from the vocal source. Levin and Lord described a procedure for dividing the analysis into octave frequency ranges, selecting pitch peaks, and determining frequency shifts. The method allowed the determination of the maximum pitch peak in each range for each time band. If the peak value exceeded a value determined by the selected threshold, then the maximum value within each range was considered to constitute a valid pitch peak. The fundamental pitch frequency shift was determined from valid peaks of adjacent time bands. Using five subjects, Levin and Lord were able to show that marked increases in the fundamental pitch frequency of the upper octave range could be used to indicate emotional change when compared to the individual's normal fundamental pitch frequency. The cepstra showed increasing fundamental pitch frequency with increasing emotional intensity.

Kuroda, Fujiwara, Okamura, and Utsuki (1976) investigated a method for determining the stress on aircraft pilots by analyzing the spectrogram of voice communications. They reported that the fundamental frequency of voiced pitch and the formants increases under stress. The elevation of the pitch of the voice, which varies directly with the tension of the vocal cords, corresponds directly to the increased emotional tension of subjects. The communication systems of aircraft are muted for frequencie; below 400 Hz to avoid interference from airframe and engine vibrations. As a result, the fundamental pitch frequency could not be reproduced from aircraft communication recordings, and it was necessary to use the sound spectrogram for analysis. The vibration space of the human voice when measured on a wide band spectrogram closely correlates to the fundamental pitch frequency of the vocal cords. This space is the vertical deflection on the spectrogram of the vowel sounds of a syllable. The vibration space shift rate (VSSR) was calculated from the vibration space of the voice during normal flight and during the emergency situation. The authors analyzed the voice communication recordings from 14 aircraft accidents, eight of them fatal. They computed the VSSR at various flight phases and compared the results to their estimates of flight stress mode (i.e., normal, urgent, and emergency). The three sets of data formed statistically separate distributions.

Simonov and Frolov (1973) investigated the use of the formant structure of the human voice for estimating the emotional stress and state of attention in aircraft pilots and cosmonauts, as well as actors. Their studies about formant structure of separate words have shown that increased emotional stress of the speaking man may be correlated with the augmentation of the mean frequency value within the range of the first formant (300-1200 Hz). They reported that, using this method, they could differentiate adequately the degree of emotional stress in 85 percent of all cases studied. They reported a good relationship between the heart rate and the changes in human voice caused by the state of emotional stress of cosmonauts during flights.

They studied the relationship between the envelope of speech and the classification of the emotional stress as positive (joy, delight) or negative (fear, anxiety). Using the outputs from five pass band octave filters (25 Hz cutoff frequency), they reported the successful estimation of changes in *emotional stress* using an empirical weighting of first and second formants, and differentiated positive from negative emotions in 90 percent of all cases with actors.

They reported that the dynamic changes in the formant maxima over time, allowed the separation of patter and slow dictation from emotional stress. The additional analysis of the dynamics of formant maxima over time, reduced the influence of voluntary and nonvoluntary distortions of speech which appear when changing its speed. The considerations of the time and sequence of the formant's place by the number and amplitude of formant maxima passages, allows the differentiation of emotional stress from accelerated or decelerated normal pronunciations.

Simonov and Frolov reported about speech signal parameters which are characteristic of the state of attention during operator activity, as compared to the state of operator rest. They used a one-third octave spectral analyzer with filter-tuning frequencies within the range of the first formant to show that the state of attention may be characterized by stabilization, a decrease in the standard deviation of the spectral components, and a drop in probability of the shift in formant maxima.

The literature review also revealed some computer algorithms which were developed to investigate various aspects of the speech signal. Rabiner and Sambur (1975) developed an algorithm for determining the endpoints of isolated utterances. Dubnowski, Schafer, and Rabiner (1976) developed an algorithm for determining the pitch of a speech signal. These two programs formed the foundation for the computer programs described below.

COMPUTER PROGRAMS

Computer programs were written in the FORTRAN programming language for use on a DEC/VAX® 11/780 digital computer, 4.0 VMS operating system. Computer software programs were developed to (1) digitize the speech analog signal (10-kHz sampling rate), (2) digitally filter the signal (900-Hz cutoff), (3) detect utterance start and end boundaries, (4) determine the pitch of voiced speech, and separate voiced, unvoiced, and noise segments of the speech, and (5) determine the formants of voiced speech. A 10-kHz sampling rate was chosen to yield a sufficient number of samples to exercise the capabilities of the programs, and the 900-Hz cutoff filter was selected to produce a relatively smooth waveform of the speech signal. The utterance- and formant-detection programs operate on the unfiltered digital signal, while the pitch-detection program operates on filtered data. The programs described below were developed by this author in accordance with the speech-processing literature as referenced. A copy of the computer programs can be made available from the author.

Utterance Detection

Speech word boundaries to define the beginning and ending of isolated words are determined with a software program generated from an algorithm originally developed by Rabiner and Sambur (1975). The program is based on a speech-versus-silence discrimination technique using energy and zero crossings for determining the boundaries of utterances. The program uses the short time average magnitude computed with a 10-ms Hamming window, and the zero crossing rate computed for a 10-ms rectangular window frame. The average magnitude is the average of the sum of the absolute amplitude values and is a measure of the energy content of the speech signal. The zero crossing rate is defined as the average of the sum of the zero crossings and is a measure of the frequency of the speech.

The program combines both energy and frequency comparisons to locate the beginning and end of speech in low signal-to-noise ratio test conditions. In the case of high fidelity recordings made in an anechoic chamber, the signal-to-noise ratio is extremely high, and the energy of the lowest level speech sounds (i.e., weak or devoiced fricatives, plosive bursts, nasals, or trailing vowels) exceeds the background noise level. A comparison of the average magnitude of the speech sounds to that of the background noise can be used to discriminate speech from the silence intervals. However, such ideal recording conditions are not often present during human factors testing, and the energy of the noise may be at the level of weak fricatives. In this case, it is difficult to locate the beginning of speech from energy comparisons alone. However, the frequency content of the speech may be radically different from that of the noise. This program attempts to overcome the potential problem of using energy alone to determine speech boundaries by including the frequency content.

The program proceeds in the following steps. First, threshold values are determined from a 100-ms header interval of ambient background noise. The mean and standard deviation of the average magnitudes and zero crossing rates for 10-ms intervals are computed. The magnitude and zero crossing rate thresholds are computed from the noise statistics and the maximum average magnitude for the header interval. Following determination of the noise statistics and function thresholds, the speech signal is processed as follows. The average magnitude is computed at 10-ms intervals to find an interval in which the values exceed a conservative threshold (calculated from the maximum magnitude of the noise in the header interval). The beginning and end points of the utterance are assumed to lie outside the 10-ms interval. Points in preceding and succeeding intervals are examined until the magnitude function exceeds a lower threshold value (also calculated from the header noise statistics). This double threshold procedure expands the word interval to tentative beginning and ending boundaries. The search in the preceding and succeeding intervals continues until the zero crossing rate falls below a threshold, again determined from the noise statistics for the header interval. The endpoints of the word boundaries are then determined and the intervals defined for which the remaining speech processing is applied. This includes determination of the rate of speed deliverance and interword spacings.

Pitch Determination

The pitch period is estimated from the magnitude of the signal and an auto-correlation function and is computed with a software program generated from an algorithm originally developed by Dubnowski, Schafer, and Rabiner (1976). The program proceeds in the following steps:

a. The speech signal, sampled at a 10-kHz rate, is digitally filtered using a finite impulse response (FIR), low pass filter with a 900-Hz cutoff frequency. The filter is symmetrical and introduces a linear phase delay of one-half the tap size to the signal.

b. The speech signal is separated into 30-ms segments (300 samples) overlapping at the ends with adjacent segments for 10 ms. In effect, the 30-ms segments are sampled every 10 ms.

c. The average magnitude computed with a 10-ms rectangular window (100 samples) for the central portion of the segment is compared to a threshold value to verify that the sample is speech and not background noise. The threshold is determined from the peak average magnitude in a 50-ms (500-sample) interval of background noise. If the average magnitude is less than the threshold, the segment is considered to be a silence interval containing background noise. However, if the magnitude exceeds the threshold, the segment is considered to be speech.

d. A clipping level is empirically determined as a fixed percentage (about 68 percent) of the least of the maximum absolute values in two adjacent 10-ms segments. The speech signal is then processed by a three-level center clipper, in which the signal is reduced to the values +1, 0, or -1, depending on whether the signal exceeds the threshold, is less than the threshold but more than the negative value of the threshold, or less than the negative value, respectively.

e. An auto-correlation function is used to determine whether the speech is voiced or unvoiced, and to calculate the pitch period if the speech segment is voiced. The auto-correlation function is computed from the clipped speech signal over the expected range of pitch periods. The largest peak of the auto-correlation function is located, and the peak value is compared to an empirically determined, fixed threshold to determine if the speech is voiced or unvoiced.

f. If the value of the highest peak exceeds the threshold, the segment is classified as voiced, and the "raw" value of the pitch period is computed from the location of the largest peak. A nonlinear smoothing operation, consisting of a 5-point median filter in series with a 3-point Hanning filter is used to smooth the pitch calculations. This filtering is applied to both the original signal and the difference between the original signal and the smoothed portion. The smoothing algorithm preserves signal discontinuities while filtering out large errors. The result of this program is a voiced/unvoiced parameter, and if voiced, the determination of the pitch.

An alternate method of pitch determination was also developed using linear predictive coding (LPC) parameters based on the simple inverse filtering tracking (SIFT) method proposed by Markel (1972). The input signal is first lowpass filtered with a 900-Hz cutoff frequency. A fourth order pole filter model is determined for the filtered signal using the auto-correlation method. The filter is sufficient to model the signal spectrum in the 0- to 1-kHz frequency range since there are only one to two formants in this range. The filtered signal is inverse filtered for the prediction error which has an approximately flat spectrum. The SIFT algorithm uses the linear predictive analysis to provide a spectrally flattened signal to facilitate pitch detection. The short time auto-correlation of the inverse filtered signal is then computed and the largest peak in the appropriate range (40 to 400 Hz) is chosen as the pitch period. An unvoiced classification is chosen when the level of the auto-correlation peak falls below a given threshold (30% of the variance).

Formant Calculation

The formants are calculated using a software program based on an auto-correlation linear predictor analysis (13th order pole, recursive filter) to represent the vocal tract (McCandless, 1974). The short-term spectrum is computed every 30 ms (300 samples), and a peak selection method based on numerical analysis, is used to find the peaks of the spectrum. The results of this program are the values of the LPC coefficients and the formant frequencies and amplitudes.

Essentially, the formants for the voiced portions of speech are computed from the smoothed spectrum which is computed using LPC analysis. The LPC coefficients are computed using the auto-correlation method with a 30-ms (300 samples) frame size Hamming window. Since the frame is indexed every 10 ms, the parameters are sampled at 100 Hz. The smoothed spectrum for the frequency response of the vocal tract response, glottal wave shape, and lips' radiation is computed from the LPC coefficients, and a peak selection routine is used to find the formant frequencies.

The predictor pole order used in this model is determined by the 10-kHz signal sampling rate. In this case, 10 resonance poles are needed to represent the vocal tract which contributes one resonance peak (one complex pole) per sampling kHz to the speech spectrum. An additional three to four poles are contributed by the source excitation and radiation load. Therefore, the total number of poles to properly represent non-nasal voiced speech is 13 to 14. An even larger number of poles is needed to represent nasals and nasalized vowels which contribute zeros to the filter transfer function (Atal & Hanauer, 1971).

The frame length of the Hamming window used in the auto-correlation method of analysis should be on the order of several fundamental pitch periods for pitch asynchronous analysis. The Hamming window is used in the auto-correlation method to taper the speech at the window ends, and the window must be long enough so that the tapering does not seriously distort the signal. A window size between 100 to 400 samples is reported to be sufficient for a 10-kHz sampling rate (Markel & Gray, 1976).

Summary Output

Software programs have been developed to present options for summarizing the speech parameters for further analysis. These are

a. A fine resolution plot of the speech signal and parameters over time. The plot shows the signals at the speech sampling rate of 10 kHz in 100-ms frames.

b. A gross resolution time plot of the signal and parameters over time. The plot shows the signals at the parameter sampling rate of 100 Hz in 10-second frames.

c. A listing of the parameter values over time. The program lists the parameter values in 10-ms increments and is a printout of the gross resolution values.

DEMONSTRATION

A DECtalk® DTC-01 synthetic speech generator (see Appendix), manufactured by the Digital Equipment Corporation (DEC), was programmed to say the same sentence under two different levels of voices. The situation represented was that of an air traffic controller giving flight instructions to an aircraft. The sentence selected was "Flight one, turn left nine degrees."

The sentence was programmed one time with the parameters that DEC considers to be the standard male voice. The parameters varied on the DECtalk® were (1) reading rate, (2) comma pause, (3) average pitch, (4) pitch range, (5) smoothness, (6) richness, and (7) laryngealization. The same sentence was then programmed in accordance with the DECtalk® operating instructions and the literature to emulate a fear-stressed voice. The reading rate was decreased to reflect longer duration of utterances, and the pitch range was increased (according to Williams & Stevens, 1972) to induce excitement into the voice (according to DECtalk®). The smoothness was increased to cause a decrease in the higher frequencies (according to DECtalk®), and the laryngealization was increased to cause irregularities in the speech at the beginning and end of the sentence and occasionally within the sentence (according to DECtalk®). The parameters programmed for each sentence are shown in Table 1.

Table 1

Voice Parameters for the DECtalk®

Parameter	Voice 1	Voice 2
Reading Rate (words per minute)	180	150
Comma Pause (ms beyond 160)	0	-10
Average Pitch (Hz)	120	160
Pitch Range (percent)	100	120
Smoothness (percent)	34	50
Richness (percent)	20	25
Laryngealization (percent)	0	70

Six people were asked to subjectively rate the two sentences for the emotional content of frustration, fear, anger, sorrow, or no emotion (neutral). Five of the six people rated the first sentence as neutral (the sixth rated it as sorrow). Three of the six people rated the second sentence as a strong fear content. The fourth rated it between fear and frustration, while the fifth and sixth rated it as a strong frustration content. As a result of this informal survey, it was decided to use these two voices as test patterns for the computer programs.

The output from the DECtalk® was fed into an analog-to-digital converter on the VAX® 11/780 for digitizing the speech signal. The digitized data were stored on a digital magnetic tapc for backup storage. The data were read from the magnetic tape to a computer disk file where the speech data for each voice were analyzed for word boundaries, pitch, and formant frequencies.

The utterance detection program was used to determine the word start and end times. The output of the program is the word boundaries located to within 10-ms increments.

The pitch determination program was used to determine the noise, unvoiced, and voiced portions of the speech, as well as the pitch for the voiced portions. The output of the program is a single parameter that identifies the speech as either noise or unvoiced or gives the pitch value for voiced portions at 10-ms intervals.

The formant calculation program was used to determine the formant frequencies of the voiced portions of the speech sample. The output of the program is the four formant values at 10-ms intervals.

The results of the analysis are shown in Figures 1 through 4 for the two voices. Figure 1 shows the word boundaries and the pitch for Voice 1. The words of the sentence are listed at the top of the figure over the corresponding boundaries. Notice that the utterance detection program did not separate the words "turn" and "left," which tend to run together in speech. In Figure 1, the smoothed pitch is shown by the solid line, while the

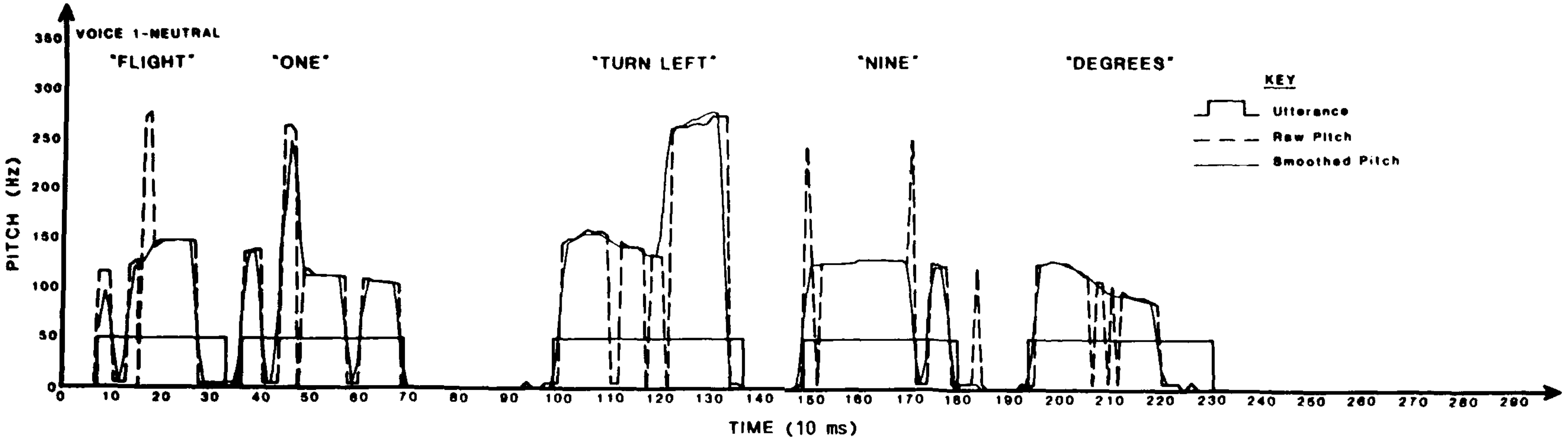

Figure 1. Pitch versus time for voice one (neutral stress).

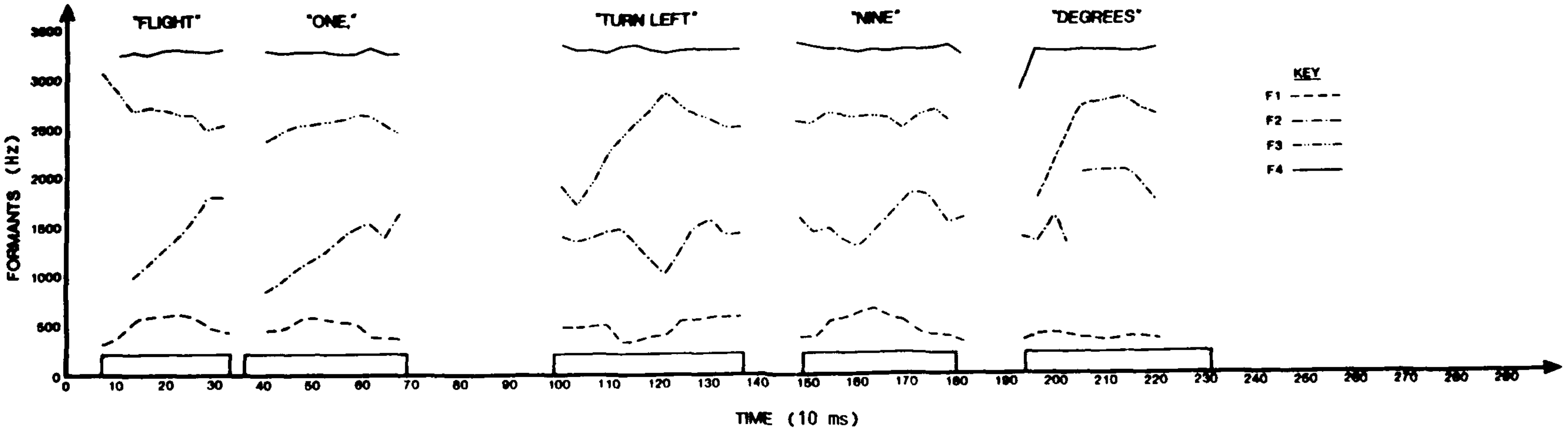

Figure 2. Formants versus time for voice one (neutral stress).

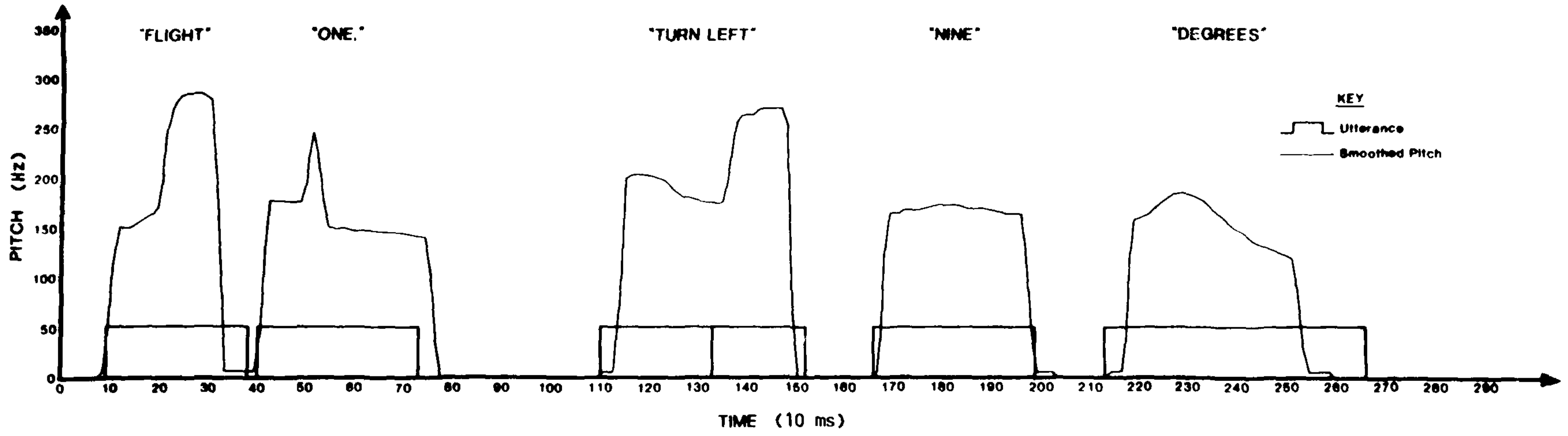

Figure 3. Pitch versus time for voice two ("fear" stressed).

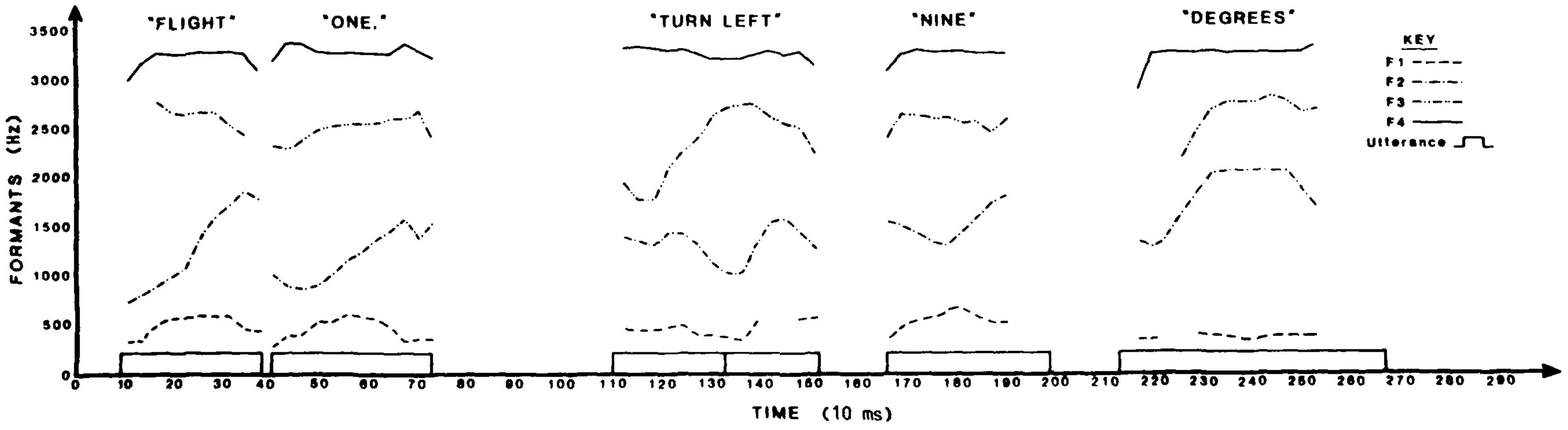

Figure 4. Formants versus time for voice two ("fear" stressed).

unsmoothed (raw) pitch estimate is shown by the dotted line for comparison. The smoothing process appears to remove much of the noise in the estimate while retaining the overall waveform. Figure 3 shows the same information (without the raw pitch estimate) for Voice 2. The smoothed pitch estimates are very similar for the two waveforms, except that the pitch for the Voice 2 is higher in magnitude.

Table 2 shows the average pitch and the standard deviation for the two speech patterns. A comparison of Tables 1 and 2 shows agreement between the programmed and computed pitch.

Table 2

Average Pitch and Standard Deviation for the Two Voices

			Pitch (Hz)
Voice	Intervals	Average	Standard Deviation
1	147	120.78	59.11
2	171	166.67	57.78

Figures 2 and 4 show the formant frequencies for the first four formants of the two voices. There are no outstanding feature differences between the two figures.

Additional sentences were input to the computer programs in which some of the parameters listed in Table 1 were varied while the remainder were held constant. The following four sentences, selected from the work of Lieberman and Michaels (1962), were used in this portion:

Sentence 1: The lamp stood on the desk.
Sentence 2: They have bought a new car.
Sentence 3: He will work hard next term.
Sentence 4: His friend came home by train.

The parameters varied and the resultant average pitch, standard deviation of the signal about the average pitch, and the number of word boundaries detected are shown in Table 3.

Table 3

Average Pitch, Standard Deviation, and Number of Word Boundaries Calculated by the Computer Programs for Various Inputs

Reading Rate:	250					
Average Pitch:	220			80		
Pitch Range:	200	100	0	200	100	0
Sentence 1	229.44	212.59	175.64	113.16	88.21	48.01
	88.60	71.18	66.20	53.23	35.34	30.92
	4	4	4	3	3	4
Sentence 2	199.70	192.31	184.34	79.46	70.32	45.58
	67.78	58.62	68.43	43.16	30.25	34.03
	2	2	2	2	2	2
Sentence 3	201.93	194.24	191.89	77.87	68.47	59.66
	63.79	66.13	57.65	47.66	31.37	27.13
	4	4	4	3	3	2
Sentence 4	219.72	204.49	189.38	100.84	95.89	37.50
	73.71	64.12	60.05	62.95	60.78	36.35
	2	3	3	3	2	3

Reading Rate:	150					
Average Pitch:	220			80		
Pitch Range:	200	100	0	200	100	0
Sentence 1	232.56	210.97	177.38	124.48	87.78	39.96
	85.97	68.36	65.98	42.50	34.84	35.22
	4	3	1	4	4	4
Sentence 2	199.80	174.97	191.99	90.56	72.90	64.45
	68.08	65.26	58.86	41.50	30.09	25.50
	2	1	2	2	2	2
Sentence 3	202.74	188.00	191.78	87.60	72.89	60.93
	63.38	66.18	57.72	32.32	27.97	24.21
	3	2	3	3	4	4
Sentence 4	221.93	202.52	190.31	103.17	71.83	48.85
	77.57	66.33	59.78	60.68	34.91	27.64
	2	3	3	3	3	3

As a further verification of the computer program's ability to correctly analyze input signals, pure tones were used as inputs. The parameters input and the calculated values are shown in Table 4.

Table 4

Pitch Determination of Pure Tones

Input		Calculated	
Avg Pitch (Hz)	Pitch Range (%)	Avg Pitch (Hz)	Std Dev (Hz)
350	0	342.33	23.13
250	0	248.19	16.79
100	0	99.27	6.74

The computed values closely agree with the pure tone generated signals.

Based on the results of the demonstration, it appears that the computer programs can distinguish changes in the voice pattern.

FUTURE GOALS AND APPLICATIONS

While the computer programs appear to distinguish voices generated by synthetic speech, it is unknown if the programs are capable of distinguishing the human voice which contains more variations. As a follow-on, it is intended to use the computer programs to analyze voice recordings of subjects using speech recognizers in voice data entry tasks under workload/stress conditions severe enough to cause misrecognitions and nonrecognitions.

If the programs continue to be successful, there are several potential benefits to be gained as follows:

a. Voice analysis may be used as an indicator of stress or operator overload. Additional analysis would be required to determine when an undesirable level of stress is reached.

b. Voice analysis may provide a range of variability that may be expected in operational environments and provide an indication of what needs to be accommodated in the development of speech recognition equipment if voice recognizers are to achieve acceptable accuracy rates.

c. Voice analysis may allow the generation of a confusion matrix which shows how the patterns of words spoken under stress changed such that they appear, to a speech recognizer, to be totally different words.

REFERENCES

Atal, B. S., & Hanauer, S. L. (1971). Speech analysis and synthesis by linear prediction of the speech wave. Journal of the Acoustical Society of America, 50, 637-655.

Dubnowski, J. J., Schafer, R. W., & Rabiner, L. R. (1976). Real time digital hardware pitch detector. IEEE Transactions on Acoustics, Speech, and Signal Proceedings, ASSP-24, pp. 2-8. Piscataway, NJ: IEEE.

French, B. A. (1983). Some effects of stress on users of a voice recognition system: A preliminary inquiry. (Thesis), Naval Postgraduate School, Monterey, CA.

Hicks, J. W., Jr. (1979). An acoustical/temporal analysis of emotional stress in speech. Doctoral dissertation, University of Florida, Gainesville, FL.

Klatt, D. H. (1980). Software for a cascade/paralleled formant synthesizer. Journal of the Acoustical Society of America, 67, 971-995.

Klatt, D. H. (1982). The Klattalk text-to-speech system. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, pp. 1589-1592. Piscataway, NJ: IEEE.

Kuroda, I., Fujiwara, O., Okamura, N., & Utsuki, N. (1976, May). Method for determining pilot stress through analysis of voice communication, Aviation, Space, and Environmental Medicine, 47, 528-533.

Levin, H., & Lord, W. (1975). Speech pitch frequency as an emotional state indicator. IEEE Transactions on Systems, Man, and Cybernetics, SMC-5(2), 259-273. Piscataway, NJ: IEEE.

Lieberman, P. (1961). Perturbations in vocal pitch. Journal of the Acoustical Society of America, 33(5), 597-603.

Lieberman, P., & Michaels, S. B. (1962). Some aspects of fundamental frequency and envelope amplitude as related to the emotional content of speech. Journal of the Acoustical Society of America, 34(7), 922-927.

Markel, J. D. (1972). The SIFT algorithm for fundamental frequency estimation. IEEE Transactions on Audio and Electroacoustics, AU-20(5), pp. 367-377. Piscataway, NJ: IEEE.

Markel, J. D., & Cray, A. H., Jr. (1976). Linear prediction of speech (1st ed). New York: Springer-Verlag.

Martin, J., & Poock, G. K. (1984). An initial look at stress and voice recognition. Journal of the American Voice Input/Output Society, 1(1), 24-33.

McCandless, S. S. (1974). An algorithm for automatic formant extraction using linear prediction spectra. IEEE Transactions on Acoustics, Speech, and Signal Proceedings, ASSP-22(2), pp. 135-141. Piscataway, NJ: IEEE.

Poock, G. K., & Martin, B. J. (1984). Effects of emotional and perceptual-motor stress on a voice recognition system's accuracy: An applied investigation. (NPS55-84-002), Monterey, CA: Naval Postgraduate School.

Rabiner, L. R., & Sambur, M. R. (1975). An algorithm for determining the endpoints of isolated utterances. Bell Systems Technology, 54(2), 297-315.

Simonov, P. V., & Frolov, M. V. (1973). Utilization of human voice for estimation of man's emotional stress and state of attention. Aerospace Medicine, 44, 256-258.

Williams, C. E., & Stevens, K. N. (1972). Emotions and speech: some acoustical correlates. Journal of the Acoustical Society of America, 52, 1238-1250.

APPENDIX

THE DECtalk® DTC-01 SYNTHETIC SPEECH GENERATOR

THE DECtalk® DTC-01 SYNTHETIC GENERATOR

The DECtalk® DTC-01 is a synthetic text-to-voice generator manufactured by the Digital Equipment Corporation. The device is programmable under computer control. Settings for voice selection, voice arguments, pause commands, stress, and syntactic structure are available. The pause commands are the length of pauses for commas and periods as well as the speaking rate. Stress includes primary and secondary lexical stress and emphatic stress. Syntactic structure control includes boundaries of a syllable; morpheme; compound noun and word; the beginning and ending of relative clauses; and the end of clauses, sentences, questions and exclamations. Seven standard voices are selectable, varying from male to female. The voice arguments which cause changes in the standard voices are (1) sex, (2) average pitch, (3) breathness, (4) forte voice, (5) head size, (6) synthesizer gain 1-5, (7) laryngealization, (8) pitch range, and (9) smoothness or high frequency attenuation. An explicit silence may be specified.

The DECtalk® was selected to generate speech for this project because of the repeatability and control of the utterances. The synthetic speech generated by the DECtalk® is extremely natural sounding. Table A-1 lists the voice parameters for the DECtalk®, and their corresponding ranges and units.

Table A-1

Voice Parameters for the DECtalk®

Parameters	Maximum	Minimum	Units
Reading Rate	350	120	words per minute
Comma Pause	250	-250	ms, beyond 160
Average Pitch	300	30	Hz
Pitch Range	250	0	percent
Smoothness	100	0	percent
Richness	100	0	percent
Laryngealization	100	0	percent

The DECtalk® is based on the work of Klatt (1980, 1982) who developed software for a cascade/parallel formant synthesizer that can generate synthetic speech on a laboratory digital computer. A flexible synthesizer configuration permitted the synthesis of sonorants by either a cascade or parallel connection of digital resonators, but frication spectra had to be synthesized by a set of resonators connected in parallel. A control program lets the user specify variable control parameter data, such as formant frequencies as a function of time points. Klatt researched strategies for the initiation of speech utterances and control parameters for the synthesis of many English sounds.